12th International Conference on Information Fusion
Seattle, WA, USA, July 6-9, 2009

Maximizing Expected Gain in Supervised Discrete Bayesian Classification When Fusing Binary Valued Features

Robert S. Lynch, Jr.
Signal Processing Branch
Naval Undersea Warfare Center
Newport, RI, U.S.A.
robert.s.lynch@navy.mil

Peter K. Willett
ECE Department
University of Connecticut
Storrs, CT, U.S.A.
willett@engr.uconn.edu

Abstract − *In this paper, previously reported work is extended for fusing binary valued features. In general, when mining discrete data to train supervised discrete Bayesian classifiers, it is often of interest to determine the best threshold setting for maximizing performance. In this work, we utilize a discrete Bayesian classification model, a gain function, to determine the best threshold setting for a given number of binary valued training data under each class. Results are demonstrated for simulated data by plotting the expected gain versus threshold settings for different numbers of training data. In general, it is shown that the expected gain reaches a maximum at a certain threshold. Further, this maximum point varies with the overall quantization of the data. Additional results are also shown for a different gain function on the decision variable, that are used to extend previously reported results.*

Keywords: Gain function, Noninformative prior, Discrete binary data, Unknown data distribution.

1 Introduction

In [11], results appeared that determined the best threshold setting for maximizing classification performance, for the problem of mining discrete data to train supervised discrete Bayesian classifiers. In this paper, it is of interest to extend this work by reporting on results when fusing binary valued features, and for using an additional gain function on the decision variable. Before elaborating on these new results, background information is provided about the methods used here (this also appeared in [11]).

1.1 Background on the Methods Used

A problem that has been well studied involves classification when the statistics (i.e., probabilistic models) of each class are unknown and determined empirically (some examples are found in [4, 3, 5, 6, 9, 13, 14, 15, 12]) from training data (i.e., supervised learning). For example, in [12] this problem was studied by showing the performance of a Bayesian classification test (referred to as the *Combined Bayes Test* (CBT)), which combines the information in discrete training and test data to infer symbol probabilities.

As previously explained in Ref. [12], by "discrete" it is meant that data used to represent each class can take on one of M possible values. This discrete data may have arisen naturally in its M-level form, or it may have been derived by quantizing "fused" feature vectors.[1] In either case, with the situation of interest there are certain labeled realizations of this (M-valued) data, and this is referred to as the "training" data under both classes. That is, in the two class case there are $N_{class\ A}$ realizations under a given class *class A* and $N_{class\ B}$ realizations under a given class *class B*. Also, given this training data, it is assumed that $N_{\mathbf{y}}$ unlabeled "test" data are observed, and these are to be simultaneously tested by a classifier. Therefore, the typical classification problem utilizing the CBT involves determining, with minimum probability of error, from which class the unknown test data have been generated.

The interesting aspect of the CBT is in its discrete observation model. In particular, the CBT was developed in [12] using the multinomial distribution for all independent discrete observations of training and test data, and the Dirichlet distribution as a noninformative prior (i.e., representing complete ignorance) on the M symbol probabilities. Basically, this implies that the prior probabilities are assumed themselves to be *uniformly distributed over the positive unit hyperplane.*

A formula for the average probability of error, $P(e)$, was also developed in [12] for the CBT, and it is typically used to illustrate its performance. In particular, based on this formula and given a fixed number of training and test data, $P(e)$ was shown to reach a minimum

[1] For example, three binary valued features can take on $M = 8$ discrete symbols: $(0,0,0)$, $(0,0,1)$, ..., $(1,1,1)$, and by the same convention four binary valued features can take on $M = 16$ discrete symbols. The point is in the data model used here the overall quantization complexity, M, can be considered to be a collection of fused features with the equivalent joint cardinality.

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
JUL 2009		**06-07-2009 to 09-07-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Maximizing Expected Gain in Supervised Discrete Bayesian Classification When Fusing Binary valued Features	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **Signal Processing Branch,Naval Undersea Warfare Center, ,Newport,RI**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM002299. Presented at the International Conference on Information Fusion (12th) (Fusion 2009). Held in Seattle, Washington, on 6-9 July 2009. U.S. Government or Federal Rights License.

14. ABSTRACT
In this paper, previously reported work is extended for fusing binary valued features. In general when mining discrete data to train supervised discrete Bayesian classifiers, it is often of interest to determine the best threshold setting for maximizing performance. In this work, we utilize a discrete Bayesian classification model, a gain function, to determine the best threshold setting for a given number of binary valued training data under each class. Results are demonstrated for simulated data by plotting the expected gain versus threshold settings for different numbers of training data. In general, it is shown that the expected gain reaches a maximum at a certain threshold. Further, this maximum point varies with the overall quantization of the data. Additional results are also shown for a different gain function on the decision variable, that are used to extend previously reported results.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Public Release**	**6**	

at a particular number of discrete symbols called M^*. Thus, the quantity M^* in the CBT is useful in that it represents the number of discrete symbols, or the joint quantization fineness of feature vectors, associated with best classification performance. [2]

1.2 New problem investigated

For the problem investigated here consider that there are N total training data, with each composed of a typical vector of fused features and an associated scalar s independent of the classification events. Also, it will be assumed that taken as aggregate s is uniformly distributed between 0 and 1. In this problem we intend to pick a threshold τ, $(0 < \tau < 1)$, such that *class A* is composed of all data having $s \leq \tau$ and *class B* has $s > \tau$. It is straightforward to see that the number of training data under *class A* is $N_{class\ A} = \tau * N$, and under *class B* it is $N_{class\ B} = (1 - \tau) * N$ (or at least the nearest integers thereto).

In modeling this problem a "gain" function, $g(s)$, is also assigned, and it is assumed that the classes and fused feature data will be represented by a trained CBT. Notice, and given $g(s)$, the expected gain obtained from "investing" in a test datum that is adjudged to be in *class B* (i.e., by the CBT) can be determined. Therefore, the problem that we address in this paper is in estimating the best threshold, τ, which yields the highest overall expected gain. This will be determined as a function of the number of data, and of the quantization complexity M representing the fused features. Further, the effect of different gain functions will also be investigated, and in this paper a new rational gain function is shown that was not investigated in [11].

2 Mathematical model for the new problem investigated

As stated above, the goal is to estimate the threshold τ that yields the highest overall expected gain when a test datum is adjudged to be in *class B*. To do this, the expected gain is defined as

$$J(\tau) = (\tau) pfa(N_{class\ A}, N_{class\ B}) E(g(s)|s \leq \tau) + (1-\tau) pd(N_{class\ A}, N_{class\ B}) E(g(s)|s > \tau) \quad (1)$$

where

τ and $(1 - \tau)$ represent prior probabilities;
$pfa(N_{class\ A}, N_{class\ B})$ is the probability of deciding *class B* in a CBT of training data of sizes $N_{class\ A}$ & $N_{class\ B}$, when in fact the test sample is truly from *class A*;
$pd(N_{class\ A}, N_{class\ B})$ is the probability of deciding *class B* in a CBT of training data sizes $N_{class\ A}$ & $N_{class\ B}$, when in fact the test sample is truly from *class B*;
From Formula (5) in Appendix A note that $pfa(N_{class\ A}, N_{class\ A}) = P(H_{class\ A})\, P(z_{class\ A} \leq \tau z_{class\ B} \mid H_{class\ A})$, and $pd(N_{class\ A}, N_{class\ B}) = 1 - P(H_{class\ B})\, P(z_{class\ A} > \tau z_{class\ B} \mid H_{class\ B})$;
The expected value of the gain function for $s \leq \tau$ is given by $E(g(s)|s \leq \tau) = \int_0^\tau \frac{1}{\tau} g(s) ds$, and for $s > \tau$ it is computed as $E(g(s)|s > \tau) = \int_\tau^1 \frac{1}{1-\tau} g(s) ds$.

In this work, to obtain results two gain functions, $g(s)$, will be utilized in Formula (1). In particular, two functions are utilized having the respective forms, $g(s) = s^c$ and $g(s) = \frac{cs}{cs+1}$.[3] Note, Figure (1) illustrates plots of the three gain functions appearing in the results below, and for comparing performance each function purposely increases with a different rate as τ is increased (i.e., into the region more favoring *class B*).

As can be seen in Formula (1) above, two integrals (i.e., for both $s \leq \tau$ and $s > \tau$) must be evaluated for each respective gain function. Table I below shows the analytical results of these integrals.

Table 1: Analytical expressions of the expected value of various gain functions for $s \leq \tau$, $E(g(s)|s \leq \tau) = \int_0^\tau \frac{1}{\tau} g(s) ds$, and for $s > \tau$, $E(g(s)|s > \tau) = \int_\tau^1 \frac{1}{1-\tau} g(s) ds$.

$g(s)$	$E(g(s)\|s \leq \tau)$	$E(g(s)\|s > \tau)$
s^c	$\frac{\tau^c}{c+1}$	$\frac{1}{1-\tau}(\frac{1}{c+1} - \frac{\tau^{c+1}}{c+1})$
$\frac{cs}{cs+1}$	$\tau - \frac{\ln(1+c\tau)}{c}$	$1 - \frac{\ln(1+c)}{c} - \tau + \frac{\ln(1+c\tau)}{c}$

3 Results

Figure (2) illustrates a plot of the average expected gain $J(\tau)$ of Formula (1), and using the quadratic gain function $g(s)s^2$ (see Figure (1), and $c = 2$ in Table 1), versus the decision threshold setting τ. In this case, four curves are shown for various quantization complexities M of respectively (top to bottom in the figure) 2, 4, 32, and 124 discrete symbols. This corresponds to, respectively, 1, 4, 5, and 7 binary valued fused features. Also, each class contains 100 samples of training data. Recall, the objective was to determine the effect that

[2] Much of the results shown in Ref. [12] was an extension of work given by Hughes, which is known in the literature as *Hughes phenomenon* (for example, see [3]).In extending Hughes' result, performance of the CBT was compared to an uncombined maximum likelihood (ML) based test. In particular, it was shown that larger numbers of test data cause M^* to increase for the CBT with an overall reduction in its average probability of error. However, for the uncombined test larger numbers of test data caused M^* to either remain unchanged or decrease, and its overall average probability of error increased. With these results, it was also shown that with a slight modification the CBT can be used to test the statistical similarity of two discrete data sets (i.e., whether they were produced by the same multinomial distribution).

[3] The variable c in both formulas is a constant.

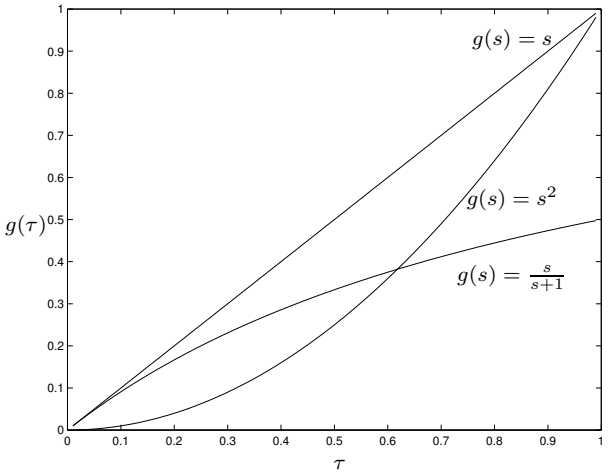

Figure 1: Illustrating a plot of the gain functions used in generating results for this paper. Shown is $g(s) = s^c$, for $c = 1$ and $c = 2$, and $g(s) = \frac{cs}{cs+1}$, for $c = 1$. Note, in each case for plotting $s = \tau$.

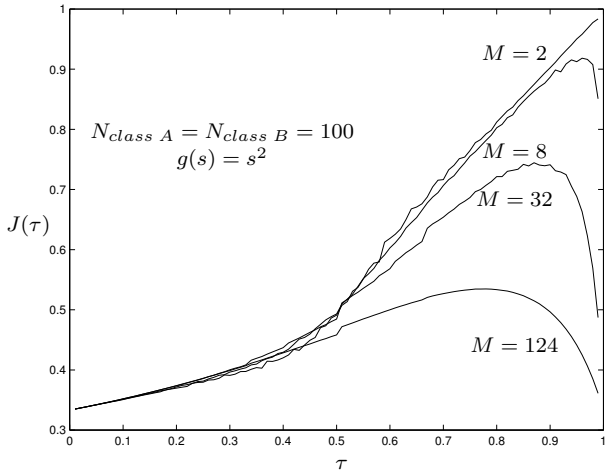

Figure 2: Illustrating a plot of the average expected gain $J(\tau)$ of Formula (1), and using the quadratic gain function $g(s) = s^2$ (see Figure (1), and $c = 2$ in Table 1), versus the decision threshold setting τ. In this case, four curves are shown for various quantization complexities M of respectively 2, 8, 32, and 124 discrete symbols.

the threshold τ has on the expected gain function of Formula (1), which can then be used to predict or estimate τ yielding best performance (i.e., when a test datum is adjudged to be in *class B*). Clearly, the expected gain reaches a maximum value in each curve (for $M = 2$, maximum $J(\tau)$is at maximum τ), and which is dependent on both τ and M. Specifically, it can be seen in Figure (2) that as M is increased from 2 discrete symbols to 124, the threshold for highest expected gain reduces from 1 to 0.75. Further, the overall absolute value of the expected gain reduces as well.

In general, notice that for the general gain function, $g(s) = s^c$, used here the average expected gain tends to increase with τ (i.e., higher gains are associated with larger threshold settings). With that, in the CBT, and for a fixed number of training data, as M is increased more uncertainty occurs in the model due to an increase in the curse of dimensionality. Thus, intuitively it is not surprising that the best overall expected gain is associated with a higher decision threshold. Further, because the curse of dimensionality predominates with larger values of M (i.e., more uncertainty in CBT cell probability estimates), it is also not surprising that the overall absolute gain decreases with M.

As a supplemental note for the results shown, all figures of this paper were obtained using Monte Carlo Simulations. Specifically, the results are based on an average of generating 50 sets of true symbol probabilities for each class (uniformly distributed), and for each of these, 100 independent trials of generating training data. Additionally, because Monte Carlo simulations were used as opposed to the complete analytical solution required in Formula (1), the results in each figure tend to have a jagged appearance.

In Figure (3), the situation of Figure (2) is repeated using instead the linear gain function $g(s) = s$ (see Figure (1), and $c = 1$ in Table 1). In this case, it can be seen that overall results are very similar to that shown in Figure (1). However, by comparing Figures (2) and (3) it can now be seen that the absolute values for the gains are larger using a linear gain function. For example, when τ is near zero in Figure (2) $J(\tau) = 0.33$, and in Figure (3) $J(\tau) = 0.33$. This implies that when operating at a best threshold setting τ, and for a given M, a linear gain function will yield the best overall average expected gain. Notice, and although not shown here, if the gain function is also scaled by a constant (e.g., in the linear case $g(s) = c * s$), then for all $c > 1$ the expected gain curves increase beyond that shown in Figure (3).

In Figure (4), the situation of Figure (2) is repeated, and using the quadratic gain function $g(s) = s^2$, but with ten samples of training data for each class. In this situation, the results have the same overall trend as in Figure (2), however, less training data has made the curves much more jagged. Further, note that the overall average gains are less, and the thresholds associated with peak gain are also less (i.e, maximum gains are shifted to the left). Further, performance results for higher values of M are more similar (e.g., compare $M = 32$ to $M = 124$). All of these trends, of course, are due to an increase in the uncertainty in the CBT model that results when very little training data is used to estimate the cell probabilities for each class.

In Figure (5), the situation of Figure (3) is repeated,

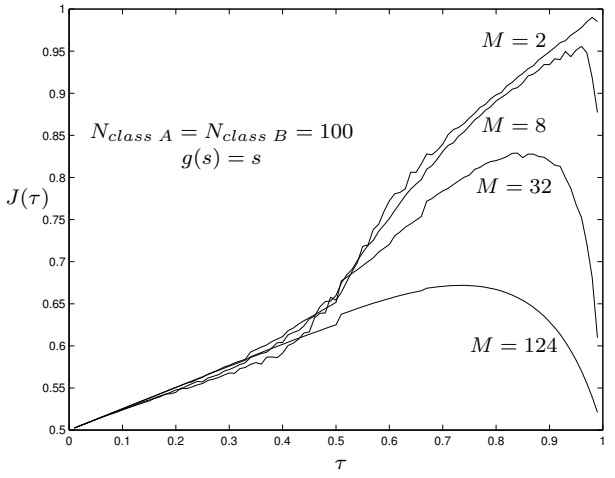

Figure 3: The situation of Figure (2) is repeated using instead the linear gain function $g(s) = s$ (see Figure (1), and $c = 1$ in Table 1).

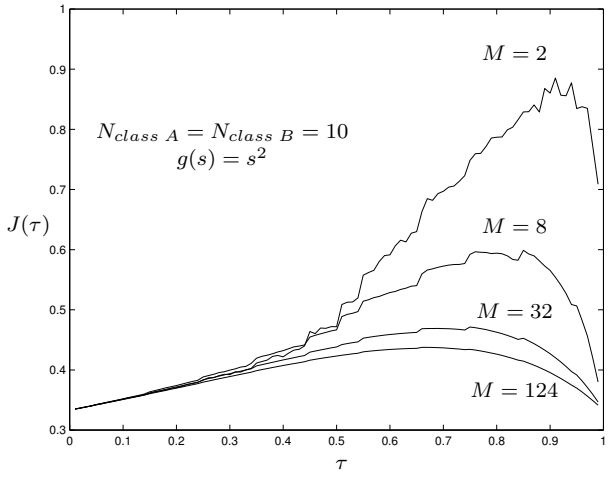

Figure 4: The situation of Figure (2) is repeated, and using the quadratic gain function $g(s) = s^2$, but with 10 samples of training data for each class.

and using the linear gain function $g(s) = s$, and again with ten samples of training data for each class. In this situation, the results have the same overall trend as in comparing Figure (4) to Figure (2) for the quadratic case above.

In Figure (6), the rational gain function $g(s) = \frac{s}{s+1}$ (see Figure (1), and $c = 1$ in Table 1) is used to obtain results, and with 100 samples of training data for each class. In this case, the rational gain function is utilized to help illustrate the importance of gain function shape on overall results. Notice in Figure (1) that both the quadratic and linear gain functions increase with increasing threshold, τ, and at a steeper rate than does the rational gain function. The impact of this on performance can be seen in Figure (6), where the overall expected gain decreases to a minimum point with τ before it finally increases again. With that, another interesting trend is that for larger values of τ maximum overall gains are now associated with larger M values (as compared to the opposite trend for either the quadratic or linear gain functions). Recall, in this case we are determining the expected gain obtained from "investing" in a test datum that is adjudged to be in *class B* (i.e., by the CBT), and that rational gain function tapers off with increasing τ. This results in a decreasing expected overall gain until τ is relatively high (i.e., *class B* has $s > \tau$), and the likelihood of data under *class B* is also very high.

In Figure (7), the situation of Figure (6) is repeated, and using the rational gain function $g(s) = \frac{s}{s+1}$, and with ten samples of training data for each class. In this situation, the results have the same overall trend as in Figure (6). However, it is also apparent that the minimum expected gain now occurs for smaller values of τ, and performance for larger values of M are more similar. As in Figures (2) through (5) this is due to the small number of training data used for estimating cell probabilities.

4 Summary

In this paper, results were demonstrated in training supervised discrete Bayesian classifiers, where it was of interest to determine the best threshold setting for maximizing expected gain in deciding on the class of an unknown test data. In this case, the CBT, and various gain functions, were utilized to determine the best threshold setting for a given number of training data under each class. Results were demonstrated for simulated data by plotting the expected gain versus threshold setting for different overall quantization levels, and for different numbers of discrete training data. In general, it was shown that the expected gain reaches a maximum at certain thresholds, which depended on the overall quantization of the data. Additionally, results were also shown for different gain functions on the decision variable. In this case, it turned out that a linear gain function produced better results than a quadratic function. Further, when using a rational gain function the expected gain actually reached a minimum point before reaching a maximum. The interesting result in this is that it the rate of increase in the gain function, with increasing decision threshold, has a large impact on the overall expected gain in correctly classifying test data.

A The Combined Bayes Test (CBT) and Its Implementation

The CBT is repeated here as it appeared in [12].

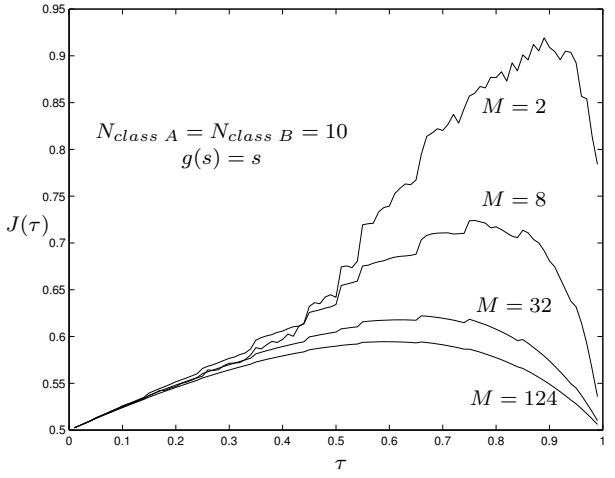

Figure 5: The situation of Figure (3) is repeated, and using the linear gain function $g(s) = s$, but with 10 samples of training data for each class.

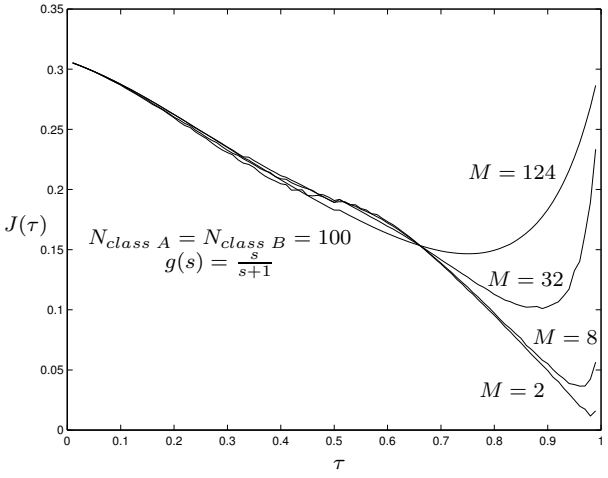

Figure 6: The situations of Figures (2) and (3) are repeated using the rational gain function $g(s) = \frac{s}{s+1}$ (see Figure (1), and $c = 1$ in Table 1), and with 100 samples of training data for each class.

A.1 Combined Information Classification

A.1.1 Combined Multinomial Model

With this model, it is assumed that there exists a pair of probability vectors, $\mathbf{p}_k$ and $\mathbf{p}_l$, the i^{th} elements of which denote the probability of a symbol of type i being observed under the respective classes k and l. The fundamental model for this testing method is thus formulated based on the number of occurrences of each discrete symbol being an i.i.d. multinomially distributed random variable. Therefore, the joint distribution for the frequency of occurrence of all training and test data with the test data, $\mathbf{y}$, a member of class k is given by (boldface indicates a vector quantity)

$$f\left(\mathbf{x}_k, \mathbf{x}_l, \mathbf{y} | \mathbf{p}_k, \mathbf{p}_l, H_k\right) = N_k! N_l! N_{\mathbf{y}}! \prod_{i=1}^{M} \frac{p_{k,i}^{x_{k,i}+y_i} p_{l,i}^{x_{l,i}}}{x_{k,i}! x_{l,i}! y_i!} \tag{2}$$

where [4]

$k, l \in \{class\ A, class\ B\}$, and $k \neq l$;
H_k is the hypothesis defined as $\mathbf{p}_{\mathbf{y}} = \mathbf{p}_k$;
M is the number of discrete symbols;
$x_{k,i}$ is the number of occurrences of the i^{th} symbol in the training data for class k;
$N_k \left\{ N_k = \sum_{i=1}^{M} x_{k,i} \right\}$ is the total number of training data for class k;
y_i is the number of occurrences of the i^{th} symbol in the test data;
$N_{\mathbf{y}} \left\{ N_{\mathbf{y}} = \sum_{i=1}^{M} y_i \right\}$ is the total number of test data;
$p_{k,i} \left\{ \sum_{i=1}^{M} p_{k,i} = 1 \right\}$ is the probability of the i^{th} symbol for class k.

[4] In the following notation k and l are exchangeable.

A.1.2 Combined Bayes Test (CBT)

Rather than assuming that $\mathbf{p}_k$ and $\mathbf{p}_l$ are simply unknown parameters to be estimated (and the resulting test a CGLRT[5]), our approach here is to give them prior distributions. Nothing a priori is known about the probability vectors, and hence the appropriate prior is one of complete ignorance: the uniform Dirichlet, which is given by

$$f\left(\mathbf{p}_k\right) = (M-1)! \mathcal{I}_{\left\{\sum_{i=1}^{M} p_{k,i} = 1\right\}} \tag{3}$$

where $\mathcal{I}_{\{x\}}$ is the indicator function.

The CBT, which can be referred to as a Bayes factor (see [7]), appears as

$$\frac{f\left(\mathbf{x}_k, \mathbf{x}_l, \mathbf{y} | H_k\right)}{f\left(\mathbf{x}_k, \mathbf{x}_l, \mathbf{y} | H_l\right)} = \frac{(N_k + M - 1)!\,(N_l + N_{\mathbf{y}} + M - 1)!}{(N_k + N_{\mathbf{y}} + M - 1)!\,(N_l + M - 1)!} \times \prod_{i=1}^{M} \frac{(x_{k,i} + y_i)!\,(x_{l,i})!}{(x_{k,i})!\,(x_{l,i} + y_i)!} \underset{H_l}{\overset{H_k}{\gtrless}} \tau \tag{4}$$

where the decision threshold τ is equal to $P\left(H_l\right) / P\left(H_k\right)$ for minimizing the probability of error.

[5] The *combined* GLRT, or CGLRT ([12]; also see, [13]), represents the correctly-posed *generalized* likelihood ratio procedure which relies on ML probability estimates culled from both training and test data. Notice that although the CGLRT is appealing from a practical perspective, from a theoretical standpoint it is less interesting due to its lack of optimality in non-asymptotic situations. With this, our preference for a Bayesian approach to this problem has been substantiated by other more recent results. Specifically, the probabilistic structure of the CBT was used in [10] with simulated and real data to reduce the number of symbols (M) for improved classification performance in a way far superior to that of GLRT based methods.

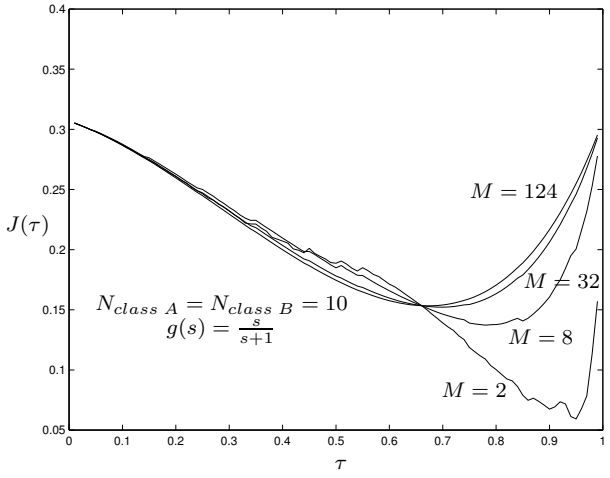

Figure 7: The situation of Figure (6) is repeated, and using the rational gain function $g(s) = \frac{s}{s+1}$, but with 10 samples of training data for each class.

Note, the CBT can be determined (after correct substitution of model parameters, and a slight reworking of the result) from the Multinomial-Dirichlet distribution shown in [1]. In fact, the data reduction method, known as the Bayesian Data Reduction Algorithm (BDRA), developed in [10] is actually based on a conditional CBT equivalent to the Multinomial-Dirichlet.

A.1.3 Probability of Error

Letting $z_k = f(\mathbf{x}_k, \mathbf{x}_l, \mathbf{y}|H_k)$ (see formula (12) above), the average probability of error for the CBT is defined as

$$P(e) = P(H_k)\,P(z_k \leq \tau z_l \mid H_k) + P(H_l)\,P(z_k > \tau z_l \mid H_l). \quad (5)$$

It is necessary to only show the first term of (5) as the second term is similar except for conditioning on H_l. Thus, ignoring $P(H_k)$, the first term of (5) is given by

$$P(z_k \leq \tau z_l \mid H_k) = \sum_{\mathbf{y}} \sum_{\mathbf{x}_k} \sum_{\mathbf{x}_l} \mathcal{I}_{\{z_k \leq \tau z_l\}} f(\mathbf{x}_k, \mathbf{x}_l, \mathbf{y}|H_k) \quad (6)$$

where $f(\mathbf{x}_k, \mathbf{x}_l, \mathbf{y}|H_k)$ was defined in formula (2) above.

References

[1] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*, Wiley, New York, 1994.

[2] L. L. Campbell, "Averaging Entropy," *IEEE Trans. on Information Theory*, vol. 41, no. 1, January 1995, pp. 338-339.

[3] K. Fukunaga, *Statistical Pattern Recognition*, Academic Press, Inc., Boston, 1990.

[4] M. Gutman, "Asymptotically Optimal Classification for Multiple Tests with Empirically Observed Statistics," *IEEE Trans. on Information Theory*, vol. 35, no. 2, March 1989, pp. 401-407.

[5] R. Hanson, J. Stutz, and P. Cheeseman, "Bayesian Classification Theory," *NASA Ames Research Center Technical Report*, no. FIA-90-12-7-01, December 1990.

[6] G. F. Hughes, "On the Mean Accuracy of Statistical Pattern Recognizers," *IEEE Trans. on Information Theory*, vol. 14, no. 1, January 1968, pp. 55-63.

[7] R. E. Kass and A. E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, June 1995, pp. 773-795.

[8] R. E. Krichevsky and V. K. Trofimov, "The Performance of Universal Encoding," *IEEE Trans. on Information Theory*, vol. 27, no. 2, March 1981, pp. 199-207.

[9] S. R. Kulkarni and O. Zeitouni, "A General Classification Rule for Probability Measures," *The Annals of Statistics*, vol. 23, no. 4, 1995, pp. 1393-1407.

[10] R. S. Lynch, Jr. and P. K. Willett, "Bayesian Classification and Feature Reduction Using Uniform Dirichlet Priors," *IEEE Transactions on Systems, Man, and Cybernetics*, **vol. 33**, no. 3, June 2003.

[11] R. S. Lynch, Jr. and P. Willett, " Estimating the Threshold for Maximizing Expected Gain in Supervised Discrete Bayesian Classification," *Proceedings of the 2009 SPIE Symposium on Security and Defense*, Orlando, FL, June 2009.

[12] R. S. Lynch, Jr. and P. Willett, "Performance Considerations for a Combined Information Classification Test Using Dirichlet Priors," *IEEE Trans. on Signal Processing*, vol. 47, no. 6, June 1999, pp. 1711-1714.

[13] N. Merhav and Y. Ephraim, "A Bayesian classification approach with application to speech recognition," *IEEE Trans. on Acoustics, Speech, and Signal Processing*, vol. 39, no. 10, October 1991, pp. 2157-2166.

[14] A. D. Wyner and J. Ziv, "Classification with Finite Memory," *IEEE Trans. on Information Theory*, vol. 42, no. 2, March 1996, pp. 337-347.

[15] J. Ziv, "On Classification with Empirically Observed Statistics and Universal Data Compression," *IEEE Trans. on Information Theory*, vol. 34, no. 2, March 1988, pp. 278-286.